A MONTE CARLO STUDY OF AGGREGATION EFFECTS ON REGRESSION PARAMETER ESTIMATES

by

Edwin Kuh and Roy E. Welsch*

Massachusetts Institute of Technology and
National Bureau of Economic Research

Working Paper No. 950-77 August 1977

A MONTE CARLO STUDY OF AGGREGATION EFFECTS ON
REGRESSION PARAMETER ESTIMATES

by

Edwin Kuh and Roy E. Welsch*

Massachusetts Institute of Technology and
National Bureau of Economic Research

Working Paper No. 950-77 August 1977

*Supported in part by NSF Grant GJ-1154X3 to the
National Bureau of Economic Research Cambridge
Computer Research Center.

ABSTRACT

The purpose of this paper is to obtain quantitative impressions, principally from Monte Carlo runs based on real data, about aggregate linear least-squares estimates in terms of their micro properties. This work builds upon theoretical studies by the same authors concerning the asymptotic properties of aggregate coefficient variances and average simple correlations among micro variables. The main results show that the Monte Carlo experiments are consistent with the theoretical results, and provide some assurance that asymptotic findings prevail in moderate sample sizes.

1. Introduction

In a previous paper [10], we studied the relationships between micro data and aggregate relations. In particular we derived several properties of coefficient variances of aggregate linear least squares estimates in terms of their micro properties. The intent of this paper is to obtain more quantitative impressions, principally from Monte Carlo runs, about aggregate estimates in terms of underlying micro characteristics. The principal results to which we shall refer subsequently are the following. First, theorem 1 of [10] is a proof that

$$(1.1) \quad [\underline{V}_N\ (\underline{b})]_{\ell\ell} \leq [BT + E + \sigma^2(N)]N(\underline{X}'\underline{X})^{-1}_{\ell\ell}$$

where B and E are constants independent of N, the number of micro components, $\sigma^2(N) = \frac{1}{N}\sum_{i=1}^{N} \sigma_{ii}$, and is assumed to be bounded, $\underline{X}$ is the TxK aggregate data matrix, $\underline{b}$ is the OLS estimate of the Kx1 aggregate parameter vector and $\underline{V}_N\ (\underline{b})_{\ell\ell}$ is the variance of the ℓ^{th} coefficient within it.

Second, it follows from this result and the boundedness of $\sigma^2(N)$ that the behavior of $N(\underline{X}'\underline{X})^{-1}$ governs the behavior of $\underline{V}_N\ (\underline{b})$. Therefore, we are interested in the equivalent expressions below which have more intuitive content,

$$(1.2) \quad N(\underline{X}'\underline{X})^{-1}_{\ell\ell} = \frac{N}{T\ S^2_\ell(1-R^2_{\ell\cdot\cdot})} = \frac{1}{Ts^2_\ell[1+(N-1)r_\ell]\ (1-R^2_{\ell\cdot\cdot})}$$

where

S^2_ℓ is the variance of macro series $\underline{X}_\ell$,

$R^2_{\ell\cdot\cdot}$ is the multiple correlation of $\underline{X}_\ell$ with the remaining K-1 explanatory macro variables,

s_ℓ^2 is the average variance among micro variables for series ℓ,

r_ℓ is the average simple correlation among micro variables for series ℓ,

T is the number of elements in each time series.

It is evident that $\underline{\underline{V}}_N(\underline{b})_{\ell\ell}$ will shrink towards zero if r_ℓ (which reflects the extent to which entities in an aggregate move synchronously) is high, s_ℓ^2 is bounded away from zero, and collinearity among macro explanatory variables reflected in $R^2_{\ell..}$ is not too great, provided that N increases. How rapidly and with what behavior that convergence takes place is one objective of this study.

Third, in circumstances where only a small amount of micro data is available as well as macro data, we derive a measure of relative efficiency, $E_{i\ell}$, which favors the aggregate when it exceeds unity.

$$(1.3)\quad E_{i\ell} = \frac{i^{th}\text{ individual micro parameter variance for variable } \ell}{\text{macro parameter variance for variable } \ell}$$

$$= \frac{[\underline{\underline{v}}_i(\underline{b})]_{\ell\ell}}{[\underline{\underline{V}}_N(\underline{b})]_{\ell\ell}}$$

$$\geq \frac{\sigma_{ii}[1+(N-1)r_\ell]}{[(\sum_k \sigma_{kk}/N)+BT+E]} \cdot \frac{\sum_{k=1}^{N} S^2_{k\ell}/N}{S^2_{i\ell}} \cdot \frac{(1-R^2_{\ell..})}{(1-R^2_{i\ell..})} .$$

The denominator of the first term on the right-hand-side of (1.3) includes parameters B and E. The former comes into play when the micro parameters are time varying, the latter when additive errors in the micro relations are correlated across individuals, thereby decreasing the relative efficiency of the macro estimates. In subsequent analysis, we have set B=E=0 so that the efficiency of using aggregates relative to micro data will be at its greatest. We have restricted ourselves to the fixed parameter case partly because the complexity added to the Monte Carlo data structure appeared not to be worth the additional cost,

and because this represents the most prevalent economic model.

Given the qualitative algebraic nature of these results, we now seek to obtain somewhat more quantitative insights about two of its aspects. First, what is the limiting behavior of macro parameter estimates, as a function of critical parameters - especially N, r_ℓ and $R^2_{\ell \cdot\cdot}$? For instance, given the latter two parameters, what is the behavior of $\underline{\underline{V}}_N(\underline{b})_{\ell\ell}$ as N increases? Second, given alternative values of the same critical set of parameters, what is the efficiency of micro compared to macro estimates?

2. Monte Carlo Structure

As a preliminary to the Monte Carlo design, we look at data on four industries for which information was available on COMPUSTAT tapes. These were used in an earlier paper [5] to study aggregation using a simple first order distributed lag investment function. Several empirical regularities influenced the Monte Carlo structure.

A. Skewness

One empirical datum of venerable standing in industrial organization literature, is the highly skewed nature of firm size distributions (analogous results hold for income distributions as well). Industrial organization literature is rife with alternative skew distributions alleged to more or less closely approximate reality, including the two parameter [1,3,4] and four parameter lognormal [9], the gamma distribution [7], and the Pareto distribution (and variants) [8]. We chose one of the more popular contenders, the two parameter lognormal [μ=0, σ^2=4] because of simplicity, widespread appeal and the fact that skewness depends directly on the coefficient of variation of the lognormal variable. As measured by shares of the five and ten largest firms, size distributions for the compustat-based industries are more skewed than the lognormal samples according to Table 1, although for Machine Tools, the shares are close. Part of the

explanation is a bias caused by the exclusion of many small and medium sized firms that are unlisted and hence, do not appear on the COMPUSTAT files. Another reason may be that the lognormal is not adequately skewed, but it should suffice for our purposes.

The significance of skewness is two-fold. It has been conjectured [5,10] that the more skewed the size distribution, the slower the decline in macro-parameter variances as N increases. Another relevant aspect is that we have presented our Monte Carlo results with increasing N for firms arrayed from largest to smallest. Of the two principal alternatives, independent random sampling or accumulation from small to large would, respectively, be unreal or foolish since, apart from an interest in smallness or representativeness as such, most information per unit about the aggregate is contained in the largest entities (barring unlikely patterns in errors of observation).

B. Synchronization

The other empirical regularity has to do with the synchronization effect designated by r_ℓ. We know because of the way in which r_ℓ enters expression (1.2) that it cannot be more than slightly negative at one extreme and, by construction, cannot exceed unity at the other. Thus for practical purposes, $1 \geq r_\ell \geq 0$. Clearly, if $r_\ell = 0$ micro series move independently of each other on the average. In these circumstances (for fixed T and s_ℓ^2), $\underline{V}_N(\underline{b})_{\ell\ell}$ will be unaffected; i.e., will not shrink according to our analysis. This formal result is in conformity with intuition, since it would be incongruous for aggregation (as distinct from averaging) to yield increased information (precision) if the elements of which it is composed are independently distributed. By the same token, maximum precision is obtained when the elements comprising the aggregate move proportionately. From data at hand for the four industries, we selected $r_\ell = .6$ as a plausible upper bound and $r_\ell = 0$ as the lower bound.

Table 2 reports average microintercorrelations for two explanatory variables, Sales and Capital Stock. One point clearly emerges. It is that the degree of synchronization is a decreasing function of sample size. Thus, while we selected $r_{\ell} = .6$ as an upper limit, that value was often exceeded for the first few firms and ordinarily wound up at half that value by the time that the full sample (ranging from 19 to 41 firms) was included.* Either larger firms are more synchronized than smaller or their behavior is just different. Whatever the truth may be, the anticipated gains in terms of parameter variance reduction when N increases will be offset by this behavior, should it exist beyond these particular samples, as we suspect to be the case.

C. Collinearity

Since the extent of correlation among macro-explanatory variables has a direct and obvious effect on macro variances, we have chosen three values of $R_{\ell..}$ to evaluate this influence in relation to other major parameters of interest. $R_{\ell..}=.98$ was chosen as the largest correlation. It is a magnitude that crops up often enough in time series to merit investigation. $R_{\ell..}=.90$ was chosen as an intermediate value, a number which, in our experience, is fairly typical for a variety of economic time series. Finally, $R_{\ell}=0$ was chosen as an unlikely but convenient extreme.

D. Number in Aggregate: N

Since we are investigating various aspects of macro parameter variances as a function of increasing N, its maximum size should be chosen with care. Cost considerations set some upper limit, while regularity or smoothness, to

*Intercorrelations for the four empirical samples were .88-.93 for Machine Tools, .98 for Petroleum Refineries, .99 for Retailers and .53-.64 for Steel; these are only weakly related to N so that full tabulation is unnecessary.

enable clearcut interpretation, sets a lower limit. After some investigation we settled on N=40, with intermediate results reported for the largest ten, twenty and thirty firms, where firm size is defined by the variance of the micro variables. Comparisons of various measures across these four groups are used to draw inferences about what happens as N increases.

E. Aggregate R^2 and Sample Size

Aggregate R^2 was somewhat arbitrarily set at .90. This magnitude is characteristic of many economic time series and in addition we wished to avoid sampling fluctuations in our results by allowing great imprecision in the macro model's goodness of fit. Similar reasoning led to a choice of T=50. That is, T=50 is a medium sized sample which helps to give a moderately high degree of precision while costs of the Monte Carlo calculations related to this aspect of the problem were not excessive.

F. Pooling

We organized the simulated micro data in three ways. The first was across time, giving N estimates of each coefficient, etc. The second method was a cross-section pool giving T estimates of each coefficient, etc. In both of these cases, the coefficient estimates were averaged to give the entries in Table 4, while the standard errors were obtained by considering the coefficient estimates as data (e.g., 10, 20, 30 or 40 data points in the time pool; 40 points in the cross-section pool) and then computing a sample standard error. The combined case pools across both time and cross section. The statistics reported are the usual ones for the single regression obtained for each level of N (10, 20, 30, 40).

G. Monte Carlo Procedure

We decided to treat the columns of the micro $\underline{\underline{X}}_i$ matrices as T independent replications of random variables, $\tilde{X}_{i\ell}$ ($i=1,\ldots,N$, $\ell=1,\ldots k$) from a kN dimensional multivariate Gaussian distribution. Our first simplification was to set k=2 and eliminate the intercept term. Thus we considered micro models with two exogenous variables and N=40.

The covariance matrix Q for $\tilde{X}_{i\ell}$ is given in Exhibit 1 with the following definitions:

$$\tilde{X}_{\ell} = \sum_{i=1}^{N} \tilde{X}_{i\ell} \qquad \ell = 1,2$$

$$S_{i\ell}^2 = \mathrm{var}\ (\tilde{X}_{i\ell})$$

$$s_{\ell}^2 = \frac{1}{N} \sum_{i=1}^{N} S_{i\ell}^2$$

$$S_{\ell}^2 = \mathrm{var}\ (\tilde{X}_{\ell})$$

$$R_{i\ell..} = \mathrm{corr}\ (\tilde{X}_{i1}, \tilde{X}_{i2}) \qquad (\text{note: } R_{i1..} = R_{i2..})$$

$$R_{\ell..} = \mathrm{corr}\ (\tilde{X}_1, \tilde{X}_2) \qquad (\text{note: } R_{1..} = R_{2..})$$

$$\alpha_{ij\ell} = \mathrm{corr}\ (\tilde{X}_{i\ell}, \tilde{X}_{j\ell})$$

$$\gamma_{ij} = \mathrm{corr}\ (\tilde{X}_{i1}, \tilde{X}_{j2})$$

$$\sigma_{ij} = \text{micro error variance}$$

$$R^2 = \text{macro multiple correlation coefficient}$$

Exhibit 1

$$Q = \begin{bmatrix} A & C' \\ C & B \end{bmatrix}$$

where

$$A = \begin{matrix} S_{11}^2 & \alpha_1 S_{11} S_{21} & \cdots & \alpha_1 S_{11} S_{N1} \\ & S_{21}^2 & & \alpha_1 S_{21} S_{N1} \\ & & \ddots & \vdots \\ \text{(symmetric)} & & & S_{N1}^2 \end{matrix}$$

$$B = \begin{matrix} S_{12}^2 & \alpha_2 S_{12} S_{22} & \cdots & \alpha_2 S_{12} S_{N2} \\ & S_{22}^2 & & \alpha_2 S_{22} S_{N2} \\ & & \ddots & \vdots \\ \text{(symmetric)} & & & S_{N2}^2 \end{matrix}$$

$$C = \begin{matrix} R_{11} S_{11} S_{12} & \gamma S_{11} S_{22} & \cdots & \gamma S_{11} S_{N2} \\ \gamma S_{21} S_{12} & R_{21} S_{21} S_{22} & & \gamma S_{21} S_{N2} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma S_{N1} S_{12} & \gamma S_{N1} S_{22} & \cdots & R_{N1} S_{N1} S_{N2} \end{matrix}$$

As the previous sections have indicated we wanted to be able to specify the distribution of the $S_{i\ell}^2$ and σ_{ii}, and the values of r_ℓ, $R_{\ell ..}$, and R^2. To do this relatively simply (and keep the computer costs in bounds) we made the following assumptions:

$$(2.1) \qquad \alpha_{ij\ell} = \alpha_\ell \text{ independent of } i \text{ and } j$$

$$(2.2) \qquad \gamma_{ij} = \gamma \text{ independent of } i \text{ and } j$$

$$(2.3) \qquad R_{i\ell ..} = R_{*\ell ..} \text{ independent of } i$$

$$(2.4) \qquad S_{i1}^2 = S_{i2}^2 \text{ all } i$$

$$(2.5) \qquad \sigma_{ii} = S_{i1}^2/100 \text{ all } i \quad .$$

In essence, 2.1 to 2.5 imply a type of interchangeability among the micro firms, except for the fact that, in general, $S_{i\ell}^2 \neq S_{j\ell}^2$, $\sigma_{ii} \neq \sigma_{jj}$ for $i \neq j$, and there are different additive error components.

By definition we have

$$r_\ell = \frac{\alpha_\ell \sum_{i \neq j} S_{i\ell} S_{j\ell}}{N(N-1)s_\ell^2}$$

or

$$(2.6) \qquad \alpha_\ell = \frac{N(N-1)s_\ell^2 r_\ell}{\sum_{i \neq j} S_{i\ell} S_{j\ell}}$$

and

$$(2.7) \qquad R_{*\ell ..} = \frac{R_{\ell ..} S_1 S_2 - \gamma \sum_{i \neq j} S_{i1} S_{j2}}{\sum_{i=1}^{N} S_{i1} S_{i2}}$$

The data generation proceeded as follows. First 1000 uniform $(0, 10^6)$ random numbers were obtained using a 31 bit linear congruential generator with a 64 cell shuffle table. These numbers were then used as seeds in generating sets of N Gaussian deviates with parameters $\mu=0$ and $\sigma^2=4$ from the Marsaglia generator [6]. The Gaussian deviates were then exponentiated to give a lognormal distribution for $S^2_{i\ell}$ $(i=1,..,N)$.

Then the α_ℓ were computed according to (2.6) and if they did not satisfy

$$(2.8) \qquad \max \ \{|\alpha_1|,|\alpha_2|\} < 1$$

another set of seeds was chosen and a new set of $S^2_{i\ell}$ generated until (2.8) was satisfied.

The next step was to find values for γ and $R_{*\ell..}$. A starting value and upper and lower bounds for γ were supplied to the program. The basic constraints on γ are that the covariance matrix (exhibit 1) we are creating must be positive definite and $|\gamma|<1$. A search procedure was developed to find a γ satisfying these constraints.

At this point T independent sets of 2N independent Gaussian variates were drawn using the Marsaglia generator (with new seeds) and then each set of size 2N was transformed using an appropriate square root of the covariance matrix, Q.

The σ_{ii} were obtained from equation (2.5) and the Marsaglia generator was again used to generate the independent Gaussian additive errors for each micro regression model.

The last step involved computing the values of β_1 and β_2 to be used in generating the micro endogenous variables. Theoretically we have

$$(2.9) \qquad R^2 = \frac{\beta^T X^T X \beta}{\beta^T X^T X \beta + T \sum_{i=1}^{N} \sigma_{ii}} .$$

We chose to set $\beta_2 \equiv 0$ so (2.9) gives

$$(2.10) \qquad \beta_1^{\,2} = \sum_{i=1}^{N} \sigma_{ii} \; \frac{R^2}{(1-R^2)S_\ell^2} .$$

The actual β_1 values were .3 for r_ℓ=0, and .06 for r_ℓ=.6. Now all of the quantities necessary to determine the micro data were available.

3. Base Case: $r_\ell=0$

As a base case, we start with $r=0$, so that no synchronization takes place and where, according to our analysis, aggregation yields few, if any, benefits.

A. Macro Parameter Behavior as a Function of N

It is clear looking at Table 3 that standard errors for both macro parameters are essentially independent of N. For given T and $R^2_{\ell..}$, $r_\ell=0$, and s^2_ℓ approximately constant, the observed independence follows directly from equation (1.2). It is equally clear that intercorrelations make a major difference and by predictable amounts. For example, with N=40 $\hat{s}_{b2}=.013$ for $R_{\ell..}=0$, $\hat{s}_{b2}=.035$ for $R_{\ell..}=.90$, and $\hat{s}_{b2}=.079$ for $R_\ell=.98$. These amounts, apart from sampling fluctuations should vary as $\sqrt{1-R^2_{\ell..}}$

A second set of questions concerns alternative uses of micro data relative to the macro data. Here we have considered three simple types of pooling, previolusly described, in full awareness that generalized least squares or maximum likelihood estimators will prove superior. These straightforware possibilities, however, are often used in actual application, perhaps after the use of appropriate covariance analysis tests for pooling. Since all the estimators are unbiased in the idealized experimental conditions deliberately chosen here (for specification problems in less favorable circumstances, see Aigner and Goldfeld [2]), we can focus attention on the standard errors alone.

Examination of just the standard errors in Table 4 indicates three clear conclusions, all of which happen to be independent of $R_{\ell..}$. First, the time series pooling yields estimated variances that closely match the macro estimates. Second, cross-sectional variability is greater than that for either the macro

or time-series pool. Since these relations do not correspond to degrees of freedom rankings, an explanation is not immediately obvious. Third, the combined cross-section, time series pool has substantially the smallest parameter standard errors of estimate; this result is not surprising since this pool has the most degrees of freedom. Unlike the other pools, estimated variances decrease as N increases. While the reasons why are not immediately evident, this last result might be dependent on the particular measure of variability that we chose.

B. Relative Efficiency

Table 5 reports four measures of efficiency - maximum, minimum, median and mean - associated with $r_{\ell}=0$ and various $R_{\ell ..}$. It should be recalled that efficiency is the ratio of the micro variance to the macro variance, so that a value exceeding unity favors the macro estimate and vice versa.

Two general results are evident. One is that there is no systematic relation to N. The other is that typical values of $E_{i\ell}$, either mean or median, are approximately unity for all three values of $R_{\ell ..}$, slightly above for $R_{\ell ..}=0$ and slightly below for $R_{\ell ..}=.98$. Since the minimum $E_{i\ell}$ for all $R_{\ell ..}$ is about .5 and the maximum is between two to three (slightly higher for $R_{\ell ..}=0$) it is quite clear from this perspective as well as that of the previous section that, under our assumptions of parameter homogeneity and $r_{\ell}=0$, there is little to choose between micro data and aggregates. Since, however, minimum $E_{i\ell}$ is about .5 and the maximum is mostly two or more, a prudent procedure would be to use the aggregate relation, but even then the anticipated benefits are not substantial.

4. Positive Synchronization: r_ℓ=.6

A. Macro Parameter Behavior as a function of N

The case of positive synchronization, it will be recalled, is one where advantages of aggregation, either in terms of coefficient variance or relative efficiency are expected to manifest themselves. There are indeed clearcut benefits in the sense of reducing the macro standard errors, as well as in terms of the efficiency measure.

Using Table 6, we see that $\hat{s}_{b1}$ for R_ℓ=.90 (and r_ℓ=.60 of course) is 0.0112 for N=10 and 0.0070 for N=40, a ratio of 1.59; ratios of similar magnitude prevail for other coefficients as well. When transformed to variances, the ratio of the variance for N=10 to that for N=40 varies between 2.11 and 2.54 (and does not interact systematically with $R_{\ell..}$). Another glance at Table 6 (this time across rows), shows that the influence of collinearity appears in the predicted manner.

Once again, two of the three simple pooling schemes provided less precise estimates than the aggregate estimates; in each instance the time series and cross-section pools had larger standard errors than the macro parameter variances, results that appear in Table 7. In this set of data with significantly positive r_ℓ, it is most interesting to compare the behavior of the complete pool standard errors with those of the aggregate. It is noteworthy that the pool parameter standard errors are about the same as those derived from the aggregate data, and decline at similar rates as N increases, even though moderate variability from parameter to parameter and experiment to experiment is apparent. Thus, in favorable circumstances such as these, aggregation provides about the same degree of estimation precision as simple pooling. Since the only benefit from generalized least squares in our context would arise from taking account of heteroscedasticity (cross-correlated errors

and serial correlation have expected values of zero), it is interesting to discover that there are possibilities for substantial estimation precision based on aggregates even when micro data are available.

B. Relative Efficiency

Unlike the situation when $r_\ell=0$, the advantages of aggregates over individual elements are strong; precision from reliance on aggregates relative to micro data is also an increasing function of N, and is adversely affected by strong collinearity among the macro explanatory variables. These aspects of the Monte Carlo runs are revealed in obvious fashion by examining Table 8.

The median efficiency measure for N=40 is about 20.5 for both $\hat{b}_1$ and $\hat{b}_2$ and $R_{\ell..}=0$. This indicator declines slightly to about 18 plus for $R_{\ell..}=.90$, but when $R_{\ell..}=.98$, median efficiency falls to about 6.5. Parallel behavior is apparent in the other efficiency measures. Since macro intercorrelations are readily observable, it is somewhat reassuring to learn that we can check for "disturbingly" high collinearity. In its absence, we have some minimal assurance that aggregation might provide some clear cut benefits over the use of small subsets of data.

5. Summary

The main results from these Monte Carlo experiments are consistent with the theoretical results in [5], and provide some assurance that asymptotic findings prevail in moderate sized samples. The first major confirmation is that the degree of synchronization, as reflected by the average micro correlation r_ℓ, exerts a highly strategic influence on macro parameters. When synchronization was nil ($r_\ell=0$) macro parameter variances did not change when the number of firms in the aggregate increased, but parameter variances fell rapidly as the number of firms increased when $r_\ell=.6$ was used. Using assumptions described at the outset, macro parameter variances based on ten firms were 60% greater than

those based on forty firms. This effect is evident in isolation and in more illuminating form when it is placed in the context of relative efficiency - a micro variance (or a measure of its central tendency) divided by the macrovariance.

The second main conclusion is that collinearity can substantially diminish otherwise beneficial effects of aggregation on macro variances. Thus, for example, our median efficiency measure is one-third of its maximum value (attained when $R_{\ell..}=0$, of course) when the correlation between macro variables reached .98. Since correlations of this magnitude are not uncommon, especially in time series applications, the well-known adverse impact of collinearity on the precision of estimation is further reinforced in connection with aggregation. While the usual qualifications about Monte Carlo studies hold in the present instance, we believe that reliance on some of the properties of four empirical subsamples increases the validity of these findings.

REFERENCES

1. Aitchison, J. and Brown, J.A.C., The Lognormal Distribution (Cambridge University Press, 1957).
2. Aigner, D.S. and Goldfeld, S., "Estimation and Prediction From Aggregate Data When Aggregates are Measured More Accurately Than Their Components," Econometrica, 42 (January, 1974), 113-134.
3. Bonini, C.P. and Simon, H.A., "The Size Distribution of Business Firms," The American Economic Review, XLVII (September, 1958), 607-617.
4. Ijiri, Yuri and Simon, H.A., "Business Firm Growth and Size," The American Economic Review, LIV (March, 1964), 77-89.
5. Kuh, Edwin, "An Essay on Aggregation Theory and Practice," in W. Sellekaerts, ed., Essays in Honor of Jan Tinbergen (New York: MacMillan, 1975).
6. Marsaglia, G., Ananthanarayanan, K. and Paul, N., "The McGill Random Number Package 'Super-Duper'," unpublished notes and program distributed at the 1972 ASA meeting in Montreal.
7. Mount, T.D. and Salem, A.B.Z., "A Convenient Descriptive Model of Income Distribution: The Gamma Density," Econometrica, 42 (November, 1974), 1115-1127.
8. Quandt, R.E., "On the Size Distribution of Firms," The American Economic Review, LVI (June, 1966), 416-432.
9. Saving, T.R., "The Four-Parameter Lognormal Diseconomies of Scale and the Size Distribution of Manufacturing Establishments," International Economic Review, 6 (January, 1965), 105-113.
10. Welsch, R.E. and Kuh, Edwin, "The Variances of Regression Coefficient Estimates Using Aggregate Data," Econometrica, 44 (March, 1976), 353-363.

TABLE 1

Cumulative Size Shares: Ranked from Large to Small Firms.

Number of Firms	COMPUTER DATA: Machine Tools	Petroleum Refineries	Retail	Steel	MONTE CARLO: Sample A	Sample B
1	.199	.258	.330	.291	.182	.143
5	.464	.596	.689	.661	.487	.384
10	.615	.819	.870	.859	.698	.551
TOTAL:	41	22	19	23	20	40

NOTE: Based on cumulative distribution of sales for the four Compustat-based industries (see text for notation and fuller explanation) and on the Monte Carlo series for which $r_\ell=0$ and $R_{\ell..}=.90$. (The A run had a total of 20 firms, and the B run 40 firms.)

TABLE 2

Average Micro Intercorrelations

MACHINE TOOLS			PETROLEUM REFINERIES			RETAIL			STEEL		
No. of Firms	Sales	Capital Stock	No. of Firms	Sales	Capital Stock	No. of Firms	Sales	Capital Stock	No. of Firms	Sales	Capital Stock
5	.61	.75	4	.68	.80	4	.54	.63	5	.67	.54
11	.50	.62	7	.58	.75	7	.44	.63	8	.65	.54
17	.48	.51	10	.47	.63	10	.40	.53	11	.55	.44
23	.43	.48	13	.43	.56	13	.37	.47	14	.49	.40
29	.40	.42	16	.38	.50	16	.34	.41	17	.43	.33
36	.37	.39	19	.35	.46	19	.30	.36	20	.40	.30
41	.33	.34	22	.32	.40				23	.36	.26

Source: Compustat-based firms and industries: see text.

TABLE 3

Estimated macro parameters and their standard errors for $r_{\ell}=0$.
(True $\beta_1=.3$, $\beta_2=0$)

N	$R_{\ell..}=.00$ $\hat{b}_1$	$\hat{s}_{b1}$	$R_{\ell..}=.90$ $\hat{b}_1$	$\hat{s}_{b1}$	$R_{\ell..}=.98$ $\hat{b}_1$	$\hat{s}_{b1}$
10	.3040	.0132	.3178	.0311	.3918	.0778
20	.2980	.0124	.2887	.0336	.3934	.0791
30	.2993	.0129	.3071	.0346	.3864	.0791
40	.3001	.0131	.3100	.0365	.3889	.0790
	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$
10	.0272	.0134	-.0366	.0302	-.1019	.0778
20	.0138	.0138	.0004	.0331	-.0963	.0791
30	.0142	.0138	-.0214	.0334	-.0907	.0790
40	.0145	.0139	-.0254	.0354	-.0935	.0788

TABLE 4

Estimated Pooled Parameters and their Standard Errors for $r_\ell=0$.

TIME SERIES POOL

N	$R_{\ell..}=.00$ $\hat{b}_1$	$\hat{s}_{b1}$	$R_{\ell..}=.90$ $\hat{b}_1$	$\hat{s}_{b1}$	$R_{\ell..}=.98$ $\hat{b}_1$	$\hat{s}_{b1}$
10	.3026	.0149	.2920	.0295	.3413	.0853
20	.2958	.0147	.2873	.0315	.3158	.0870
30	.2974	.0135	.2960	.0310	.3200	.0785
40	.2994	.0143	.2979	.0340	.3084	.0769
	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$
10	.0002	.0149	.0063	.0194	-.0421	.0927
20	-.0008	.0143	.0074	.0235	-.0175	.0880
30	-.0003	.0135	.0003	.0253	-.0220	.0802
40	-.0013	.0140	-.0018	.0285	-.0102	.0773

CROSS SECTION POOL

N	$R_{\ell..}=.00$ $\hat{b}_1$	$\hat{s}_{b1}$	$R_{\ell..}=.90$ $\hat{b}_1$	$\hat{s}_{b1}$	$R_{\ell..}=.98$ $\hat{b}_1$	$\hat{s}_{b1}$
10	.3089	.0590	.2986	.1233	.3373	.4080
20	.3045	.0448	.2993	.0756	.3501	.3394
30	.3042	.0424	.3004	.0688	.3553	.3247
40	.3043	.0411	.3002	.0676	.3558	.3206
	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$
10	-.0021	.0579	.0001	.1286	-.0308	.4057
20	-.0002	.0467	-.0031	.0843	-.0444	.3353
30	.0000	.0443	-.0035	.0752	-.0497	.3207
40	.0000	.0433	-.0034	.0738	-.0504	.3166

COMBINED POOL

N	$R_{\ell..}=.00$ $\hat{b}_1$	$\hat{s}_{b1}$	$R_{\ell..}=.90$ $\hat{b}_1$	$\hat{s}_{b1}$	$R_{\ell..}=.98$ $\hat{b}_1$	$\hat{s}_{b1}$
10	.3083	.0051	.3041	.0099	.3893	.0244
20	.3061	.0035	.2994	.0071	.3831	.0172
30	.3060	.0029	.3004	.0058	.3826	.0140
40	.3060	.0025	.3002	.0050	.3822	.0121
	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$
10	.0066	.0050	-.0079	.0095	-.0894	.0242
20	.0056	.0035	-.0045	.0069	-.0833	.0171
30	.0055	.0028	-.0053	.0056	-.0828	.0139
40	.0054	.0024	-.0052	.0049	-.0824	.0120

TABLE 5

Relative Efficiency Measures for $r_\ell = 0$.

$R_\ell = .00$

N	MAXIMUM	MINIMUM	MEDIAN	MEAN
		$\hat{b}_1$		
10	2.79	0.68	1.09	1.28
20	3.13	0.76	1.27	1.43
30	2.93	0.71	1.24	1.33
40	2.84	0.69	1.23	1.33
		$\hat{b}_2$		
10	2.58	0.47	1.16	1.25
20	2.40	0.44	1.11	1.22
30	2.42	0.44	1.11	1.17
40	2.37	0.43	1.08	1.14

$R_\ell = .90$

N	MAXIMUM	MINIMUM	MEDIAN	MEAN
		$\hat{b}_1$		
10	1.59	0.68	1.04	1.12
20	1.87	0.58	0.97	1.01
30	1.84	0.52	0.90	0.93
40	1.65	0.47	0.83	0.87
		$\hat{b}_2$		
10	1.66	0.58	1.20	1.17
20	1.71	0.48	1.06	1.03
30	2.19	0.47	1.03	1.01
40	1.95	0.42	0.91	0.92

$R_\ell = .98$

N	MAXIMUM	MINIMUM	MEDIAN	MEAN
		$\hat{b}_1$		
10	2.26	0.51	0.84	1.08
20	2.19	0.50	0.85	0.94
30	2.19	0.50	0.85	0.93
40	2.19	0.44	0.83	0.89
		$\hat{b}_2$		
10	2.14	0.53	0.90	1.09
20	2.06	0.51	0.85	0.94
30	2.07	0.51	0.86	0.92
40	2.08	0.43	0.84	0.88

TABLE 6

Estimated Macro Parameters and their Standard Errors for $r_{\ell}=.60$.
(True $\beta_1=.06$, $\beta_2=0$)

	$R_{\ell..}=.00$		$R_{\ell..}=.90$		$R_{\ell..}=.98$	
N	$\hat{b}_1$	$\hat{s}_{b1}$	$\hat{b}_1$	$\hat{s}_{b1}$	$\hat{b}_1$	$\hat{s}_{b1}$
10	.0581	.0051	.0659	.0112	.0597	.0206
20	.0583	.0041	.0668	.0083	.0745	.0173
30	.0574	.0038	.0671	.0074	.0753	.0148
40	.0576	.0035	.0646	.0070	.0747	.0140
	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$
10	.0078	.0054	-.0099	.0125	.0045	.0205
20	.0083	.0044	-.0088	.0092	-.0104	.0174
30	.0079	.0041	-.0088	.0082	-.0127	.0149
40	.0078	.0037	-.0062	.0077	-.0120	.0140

TABLE 7

Estimated Pooled Parameters and their Standard Errors for r_{ℓ}=.60.

TIME SERIES POOL

	$R_{\ell..}$=.00		$R_{\ell..}$=.90		$R_{\ell..}$=.98	
	$\hat{b}_1$	$\hat{s}_{b1}$	$\hat{b}_1$	$\hat{s}_{b1}$	$\hat{b}_1$	$\hat{s}_{b1}$
10	.0607	.0144	.0605	.0236	.0754	.0346
20	.0587	.0164	.0631	.0216	.0704	.0313
30	.0562	.0161	.0621	.0223	.0721	.0332
40	.0568	.0158	.0547	.0290	.0700	.0314
	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$
10	.0105	.0149	-.0048	.0309	-.0107	.0373
20	.0080	.0207	-.0038	.0249	-.0066	.0340
30	.0076	.0211	-.0026	.0260	-.0117	.0369
40	.0070	.0192	.0050	.0313	-.0079	.0335

CROSS SECTION POOL

	$R_{\ell..}$=.00		$R_{\ell..}$=.90		$R_{\ell..}$=.98	
	$\hat{b}_1$	$\hat{s}_{b1}$	$\hat{b}_1$	$\hat{s}_{b1}$	$\hat{b}_1$	$\hat{s}_{b1}$
10	.0782	.2504	.0805	.2111	.0612	.1722
20	.0578	.1693	.0878	.1435	.0661	.1023
30	.0624	.1456	.0815	.1231	.0739	.0830
40	.0621	.1404	.0801	.1159	.0748	.0756
	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$
10	.0228	.2207	.0191	.2606	.0120	.1651
20	.0238	.1478	.0004	.1585	.0028	.1009
30	.0255	.1308	.0000	.1466	.0006	.0843
40	.0243	.1267	.0009	.1408	.0009	.0798

COMBINED POOL

	$R_{\ell..}$=.00		$R_{\ell..}$=.90		$R_{\ell..}$=.98	
	$\hat{b}_1$	$\hat{s}_{b1}$	$\hat{b}_1$	$\hat{s}_{b1}$	$\hat{b}_1$	$\hat{s}_{b1}$
10	.0567	.0049	.0664	.0093	.0804	.0112
20	.0572	.0035	.0660	.0067	.0781	.0078
30	.0569	.0028	.0657	.0054	.0786	.0064
40	.0569	.0025	.0654	.0047	.0785	.0055
	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$	$\hat{b}_2$	$\hat{s}_{b2}$
10	.0080	.0052	-.0095	.0105	-.0146	.0111
20	.0088	.0037	-.0083	.0075	-.0134	.0078
30	.0086	.0030	-.0082	.0061	-.0145	.0064
40	.0085	.0026	-.0080	.0052	-.0144	.0055

TABLE 8

Relative Efficiency Measures for r_ℓ=.60.

R_ℓ=.00

	MAXIMUM	MINIMUM	MEDIAN	MEAN
		$\hat{b}_1$		
10	13.78	6.90	9.23	9.75
20	21.11	10.48	14.83	15.29
30	25.99	9.73	16.96	17.19
40	31.29	11.71	20.41	20.48
		$\hat{b}_2$		
10	13.94	6.18	9.45	9.79
20	22.14	9.42	15.20	15.30
30	25.22	9.49	17.03	17.16
40	30.39	11.43	20.53	20.40

R_ℓ=.90

	MAXIMUM	MINIMUM	MEDIAN	MEAN
		$\hat{b}_1$		
10	9.41	4.65	7.18	6.87
20	22.42	8.39	13.36	13.53
30	28.48	9.84	16.71	16.66
40	37.42	10.90	18.93	19.17
		$\hat{b}_2$		
10	10.26	4.08	6.94	6.80
20	24.25	7.47	12.81	13.53
30	30.98	9.03	16.31	16.42
40	37.62	10.03	18.22	18.78

R_ℓ=.98

	MAXIMUM	MINIMUM	MEDIAN	MEAN
		$\hat{b}_1$		
10	4.06	2.57	2.91	3.10
20	7.05	2.44	4.15	4.36
30	10.27	3.33	5.99	6.24
40	11.55	3.41	6.44	6.83
		$\hat{b}_2$		
10	3.97	2.27	3.15	3.11
20	6.95	2.31	4.32	4.39
30	10.28	3.14	6.18	6.31
40	11.59	3.54	6.76	6.85

(#946 + 947 not used)